AD_________________

GRANT NO: DAMD17-94-J-4345

TITLE: Evaluation of Digital Mammography Display

PRINCIPAL INVESTIGATOR(S): Etta Pisano, M.D.

CONTRACTING ORGANIZATION: University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-4100

DTIC
ELECTE
NOV 2 1 1995
S D
G

REPORT DATE: September 1995

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

19951116 086

DTIC QUALITY INSPECTED 5

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1995	Annual 1 Sep 94 - 31 Aug 95

4. TITLE AND SUBTITLE	5. FUNDING NUMBERS
Evaluation of Digital Mammography Display	DAMD17-94-J-4345
6. AUTHOR(S) Etta Pisano, M.D.	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of North Carolina at Chapel Hill Chapel Hill, North Carolina 27599-4100	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012	10. SPONSORING/MONITORING AGENCY REPORT NUMBER
11. SUPPLEMENTARY NOTES	
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited	12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

The purpose of this research is to experimentally determine the diagnostic accuracy and interpretation speed of digitally acquired mammograms displayed on the best available display methods.

We propose to conduct an ROC study comparing a film based display to the best available state-of-the-art electronic workstation.

During the first year we have carried out experiments to determine the parameter values to be used for intensity windowing applied to mammograms. For both calcifications and spiculations, we found statistically significant improvement in detection with specified values for the intensity windows [Pisano '95]. These results will be incorporated into the design of the clinical ROC experiment where the video monitors are the display devices.

We have developed a computer model of mammography interpretation based on eyetracking studies completed during this last year [Beard '95]. The model allows complex tasks to be graphically evaluated and thereby allow the rapid comparison of the image manipulation time of many design alternatives. We believe the time required to manipulate images will be the most important factor in selecting a workstation design. The initial development of the mammography workstation is underway and should be completed during year 2 in time for the clinical ROC studies to begin.

14. SUBJECT TERMS	15. NUMBER OF PAGES
DIGITAL MAMMOGRAPHY, CLAHE, WORKSTATIONS, MAMMOGRAPHY	52
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the US Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

EP For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

[signature] 9/26/[illegible]

PI - Signature Date

TABLE OF CONTENTS

Accesion For

NTIS CRA&I ☒
DTIC TAB ☐
Unannounced ☐
Justification

By
Distribution /

Availability Codes

Dist	Avail and / or Special
A-1	

Introduction

A. Nature of the problem (from original proposal)

A new type of digital mammography device has been developed at the University of Toronto. This scanning slot digital mammography system provides 50um, 12-bit pixels with inherently better contrast than that of conventional mammogram. The advent of digitally acquired mammograms offers the possibility of further improvements in early breast cancer detection. Specifically, digital acquisition systems decouple the process of x-ray photon detection from image display by using a primary detector that directly quantifies transmitted photons. This allows digital systems to be more efficient in utilization of radiation dose. Digital systems also allow a wide dynamic range so that a wider range of tissue contrast can be appreciated. Subtle contrast differences can be amplified and the distinction between benign and malignant might be increased. The new Toronto scanning slot digital mammography system has the further advantage of reduced scatter compared with both conventional and phosphor plate technologies. Furthermore, digital systems have the capacity to bring revolutionary advantages to breast cancer detection and management: 1) image processing for increased lesion conspicuity; 2) computer-aided diagnosis for enhanced radiologic interpretation; 3) teleradiology, or image transmission, as a means of bringing world-class expertise to community hospitals and remote areas; 4) improved image access and communication through digital image archiving and transmission; and 5) dynamic, or "real time" imaging for use during biopsy and localization procedures.

However, there are limitations to both laser-printed film and electronic displays, the two possible display methods for digital mammography. The best quality film printers can only display 87um pixels in an 8"X10" printing of the digital data. This would not provide sufficient spatial bandwidth for the available data. These printers may also lack sufficient greyscale bandwidth. The best possible 2500x2000 pixel monitors can generate over 170-680 nits luminance without pixel bloom. To gain access to the full grey scale bandwidth, monitor display would require intensity windowing, and to view the image at the full 50 mm spatial resolution, roaming and zooming would be necessary. Clearly, any display modality requires compromises that will effect diagnostic accuracy and interpretation speed.

B. Background of previous work (from original proposal)

For a number of years, the Medical Image Presentation research group at UNC-CH has been exploring various issues concerning the display of medical images. Early on we addressed the issues of standardization of display devices to assure legitimate comparison of various display methods under investigation. The display is perceptually linearized so that each intensity step in the acquired image is displayed as an equally perceptible step in the grey levels of the display [Pizer 1981, 1987, 1989, Johnston 1985, Rogers 1987]. In addition, our group, under another grant, (RO1 CA44060) has developed and experimentally evaluated the ergonomic and cognitive aspects of

electronic workstations. We constructed a prototype workstation called FilmStrip using a single 2048x2560 pixel high-brightness monitor, a very simple interaction, and an extremely fast image display time (0.1 sec). A controlled subject experiment was used to evaluate FilmStrip relative to film and alternator [Beard 1993]. All reports were of clinically acceptable accuracy. Based on our experimental results, we are 95% confident that FilmStrip is no more than 1.5 minutes faster and no more than 30 seconds slower than film. This is the first time a radiology workstation has been shown to be as fast as film for interpretation of medical images under clinically realistic conditions. We have conducted a subsequent experiment showing that a lower cost version of FilmStrip called FilmStriplet can also be clinically viable with sufficient training [Beard 1993].

Under a medical image presentation program project grant, (P01-CA47982), we have been exploring different image processing methods, specifically various versions of the Contrast Limited Adaptive Histogram Equalization algorithm, and have developed an experimental method to optimize the parameters for a given enhancement algorithm that takes into account the deleterious effects of image noise and that does not require the performance of a full clinical trial [Puff, 1992]. This work has involved the conduct of a number of image quality assessment experiments.

Under the previously described interactive Digital Mammography Development Group grant, Gray Scale Image Processing For Digital Mammography, (R01 CA 60193), we are conducting preliminary experiments to determine the effect of the variable amount of radiographically dense breast tissue, the mammographic characteristics of various lesion types, and the location of lesions within the breast on the choice of appropriate intensity windows and other image processing algorithms selected for electronic viewing of mammograms. The results of this investigation will also give us some indication of the number of intensity windows that might be useful, or needed, for display of the recorded digital information.

C. Purpose of present work.

The purpose of this study is to determine experimentally the diagnostic accuracy and interpretation speed of the available display methods.

D. Methods of approach

We propose to conduct an ROC study involving the best available display methods, one representative of a film based display, and one using the best available state-of-the-art electronic workstation.

Research Methods and Results to date:

1. To achieve the goals of this research, we propose using digitally acquired mammograms. At this point in time the availability of the clinical digital units have been delayed until sometime in the fall of this year, '95 to early '96. Conduct of the actual ROC observer studies has therefore been delayed until the clinical images become available.

2. Since the inception of this grant, a number of technical advances have been made that directly modify the experimental procedures to be carried out under this proposal. A major change is that there are now laser printers that can meet the requirements for display of mammograms on an 8"x10" format with 12 bits of gray levels.

We are obtaining such a laser printer (Kodak) from internal funds along with a Fischer Digital Mammography unit to be located at UNC Hospitals also from internal funding. The presence of this unit along with the digital images to be obtained from Thomas Jefferson hospital will provide us with more digital mammograms than originally expected. Thus, the delay in obtaining the digital units is offset by the eventual increased availability of clinical images. With the availability of the new laser printer, we no longer need to optically reduce larger format laser images.

3. During the first funding period of this grant, a number of changes in the state-of-the-art of monitor technology has occurred, a) the original high-brightness monitor that was promised during the inception of this grant was not developed by the manufacturer. However, over the same period of time, several manufacturers have made available high-brightness monitors ranging in maximum luminance from 150 ftL. to 200 ftL.

Unfortunately, the interface electronics to drive 2k x 2.5k monitors from conventional host computers has lagged behind and only now are becoming available. We have placed an order with TechSource for their system which will be capable of driving up to 4 such monitors (necessary for a realistic clinical evaluation). A prototype of this system will be delivered within the next month. This will enable us to begin installation and further development of the mammography workstation software in preparation for the clinical studies. With the assistance of Dr. Beard, Mr. Hemminger and two graduate students, we have started the workstation design.

4. Workstation development. We have developed a QGOMS model of mammography readings based on eyetracking studies completed during this last year [Beard '95] allows users to model complex tasks graphically. This tool will allow the rapid comparison of the image manipulation time of many design alternatives. We believe the time required to manipulate images will be the most important factor in selecting a workstation design. The initial development of the mammography workstation is underway and should be completed during year 2 in time for the clinical ROC studies to begin.

5. During the last year we have carried out experiments to determine the parameter values to be used in conducting observer experiments to evaluate the use of intensity windowing and contrast limited adaptive histogram equalization (CLAHE) applied to mammograms. We completed observer studies using CLAHE, and found significant improvement in the detection of spiculations [Garrett 95]. We also completed observer studies with preset intensity windows selected for spiculations and calcifications. Our results showed statistically significant improvement in detection of both features with

specified values for the intensity windows [Pisano '95]. These results will be incorporated into the design of the clinical ROC experiment where the video monitors are the display devices. This research is also partially supported by NIH R01-CA60193.

6. Since one of the two display devices will be laser printed films, we spent effort investigating the characteristics and the variables that must be controlled or understood when printing digital images onto film with the laser printer. Although we will be using the Kodak printer for the clinical mammograms, we were required to develop the techniques and gain experience with our Lumisys laser printer to carry out the intensity window observer studies [see appendix].

We have also implemented perceptual linearization of both laser printer and video monitor display systems in collaboration with the proposed ACR/NEMA standards working committee.

Conclusions

Although the acquisition of digital mammograms has been delayed by a factor of about 6 months, we have made significant progress in:

1. Evaluation of the intensity windowing as an image enhancement method,
2. Developing the methods for and identifying the critical areas of quality control for the laser printed images.
3. Evaluating the transfer characteristics of the laser printer and the video monitors.
4. Developing the software tools for the electronic mammographic workstation.

Proposed research for the 02 year period:

1. Complete the software development of the electronic mammography workstation.
2. Identify and purchase the best available high brightness and high resolution video monitors and associated electronics. The funding for the workstation is partly from this grant and partly from R01- CA60193.
3. To install the Fischer digital mammographic unit and Kodak laser printer into UNC Hospitals. To begin the acquisition of clinical data which will be available to this project for evaluation of the display methods.
4. To redesign the experimental protocol for improved and more efficient data collection to meet the goals of this grant. The redesign in no way alters the ultimate goal of this research. Primarily, it accommodates the advances in technology that has occurred since the original experiments were proposed and should result in improved ROC studies.
5. As a result of the delay in availability of clinical digital mammograms, we have operated under a reduced budget during the 01 year, and propose to operate under a reduced budget during part of the 02 year until the clinical images are being obtained.

References:

Beard DV, Hemminger BM, Perry JR, Mauro M, Muller K, Warshauer D, Zito A, and Smith M Single-Screen Workstation vs. Film Alternator for fast CT Interpretation, *Radiology*, 1993; 187(2):1-6.

Johnston RE, Zimmerman JB, & Pizer SM: "Perceptual standardization", Proc. SPIE 536, 44-49, 1985.

Pizer SM Intensity mappings to linearize display devices, Comp. Graph. Image Proc.17, 262-268, 1981.

Pizer SM, Rogers D, Johnston RE, & Beard, DV "Effective Presentation of Medical Images on an Electronic Display Station", *RadioGraphics*, Vol. 7, No. 6, 1267-1274, Nov. 1987.

Pizer SM, & Beard DV. "Medical Image Workstation: State of Science & Technology". *Journal of Digital Imaging*, Nov. 1989 2(4) 185-193.

Rogers C, Johnson RE, Hemminger BM, Pizer SM, (1987) Effect of Ambient Light on Electronically Displayed Medical Images as Measured by Luminance Discrimination Thresholds. *J. Optical Society Am* 4(5) 976-983.

Beard DV, Bream P, Pisano ED, Conroy P, Johnston RE, Braeuning P, McLelland R, Clark R. "Eye Movement During Mammography Interpretation: Eyetracker Results and Workstation Design Implications. Submitted to *Academic Radiology*. (See Appendix A.)

Garrett WF, Pisano ED, Puff DT. "Does CLAHE Improve Detection of Spiculations in Digitized Mammograms?". Manuscript in preparation (See Appendix B.)

Pisano ED, Chandramouli J, Hemminger BM, Johnston RE, Muller KE, Pizer SM. "The Utility of Intensity Windowing in Improved Detection of Simulated Masses in Mammograms of Dense Breasts". To be Presented at RSNA, 1995.

APPENDIX A

Eye Movement During Mammography Interpretation : Eyetracker Results and Workstation Design Implications.

David V Beard, Ph.D[1,2]., Peter Bream[1] , M.D., Etta D Pisano[1], M.D., Pat Conroy[2], B.S., R Eugene Johnston[1], Ph.D. Patricia Braeuning[1], M.D., Robert McLelland[1], M.D., Richard Clark[1], M.D.

1. Department of Radiology and 2, Department of Computer Science
University of North Carolina,
Chapel Hill, NC 27599-7510

After August 31, David Beard will be at The State University of Idaho
(919) 966-5467 beard@cs.unc.edu

This research was in part supported under NIH research grant R01 CA 60193 and DOD grant DAMD17-94-J-4345.

Correspondance should be with R. Eugene Johnston, Ph.D.
Department of Radiology
University of North Carolina
Chapel Hill, N.C. 27599-7510

Phone (919) 966-5069, FAX (919) 966-5934

Abstract

Rationale and Objectives: Digital mammography can potentially improve mammography image and interpretation quality. On-line interpretation of these images from a workstation may improve interpretation logistics and increase availability of comparison images. Workstation interpretation of eight 4x5k pixel mammograms on two or four 2.5x2k pixel monitors is problematic due to the time spent in choosing which images to display on which monitors and zooming and roaming on individual images that are too large to display at full resolution.
Methods: We used an eyetracker to measure radiologists viewing behavior during mammography interpretation.
Results: A significant portion of the mammographers time is spent viewing "comparison pairs" such as the left mediolateral (MLO) and cranio-caudal (CC) or the old and new left cranio-caudal images.
Conclusions: We estimate the number of required image display, zoom, and roam operations as a function of the number of monitors for a potential mammography workstation. From time-motion analysis we can predict the viability of mammographic workstations.

Keywords. Eyetracking, Digital mammography, image display

1. Introduction

Screening mammography is an effective procedure for early identification of breast cancer [1-10]. Mammography imaging technology has improved significantly in the last 20 years including the development of dedicated mammography equipment with appropriate x-ray beam quality, grid capability, adequate breast compression, automatic exposure control, better film screen systems, and appropriate film processing [11,12]. Nevertheless, roughly 10% of clinically obvious breast cancers are not visible with mammography [4], most frequently in patients with large amounts of breast glandular tissue [4, 1]. Further, near optimal film processing is critical [14], and film-based mammography is often inaccessible in rural locations with insufficient population to justify a proximately-located mammographer.

Digital mammography has the potential to alleviate some of these problems [15]. Typically, such systems generate a 4000x5000 12bit/pixel matrix for each image in the mammography study. Preliminary evaluation of scanning slot approaches indicates enhanced greyscale resolution over film-screen mammography [16] which may improve detection under conditions of large amounts of breast glandular tissue. Digital mammography would also allow film-less interpretation and teleradiology to remote locations.

However, display of digital mammography is problematic. Current film printers can only print the 4000x5000 pixel matrix if physically larger-than-normal films are used, which would generate ergonomic difficulties during film/alternator interpretation. However, there is a new generation of printers becoming available that can print on an 8" x 10" format at 50 μm/pixel. It is possible that even with printers, intensity windowing, or some other greyscale manipulation approach may be needed to best present the dynamic range of the grey scale data, Finally, film development and handling are logistically troublesome. A mammography workstation that facilitates fast and accurate on-line interpretations would be of immense value to mammography clinics.

Monitor quality has improved significantly over the last several years with the current best quality 70hz monitors generating 150fL of luminance and displaying a 2500x2000 pixel image in as little as 0.11 seconds. Although some further increases in luminance can be expected, monitors are not likely to produce the high brightness of a film lightbox. Monitors can be improved in their noise characteristics and thereby improve dynamic range. It is still unlikely that

future monitors will have sufficient greyscale dynamic range to allow interpretation without intensity windowing or other greyscale filtering.

Simply using eight or more of these monitors is not likely to produce a viable workstation. Such high performance monitors are typically larger than 8"x10" mammography film. The large physical size of the resulting workstation would be prohibitive in many space conscious clinics, and would require considerable time for the mammographer to move physically back and forth while trying to compare various images. Further, these monitors are likely to be very expensive making the cost of an eight-monitor workstation prohibitive.

Thus, two significant ergonomic obstacles remain. First, since only two or four monitors can be realistically used in a financially viable workstation, the radiologist needs to constantly chose which images are to be displayed on which monitor. Second, the mammographer must roam and zoom over a 4000x5000 pixel image to see it at full resolution on a 2500x2000 pixel monitor. Both roam and zoom, and image-display selection are cognitively complex tasks, disrupting the mammographer's concentration during interpretation. These tasks will require many time-consuming hand motions and button presses, as well as time to wait for the system to display images, all of which can add up to an additional two to four minutes of radiologist time, while an interpretation on film would require less than a minute.

Thus, answers are needed to a number of critical questions that can significantly effect the viability of the mammography workstation concept. How often do mammographers need to roam around the full resolution image, and how often can they manage with a lower resolution image? How often will mammographers want to change which images are being displayed? Which sequence of images will they choose to display next? How fast must a monitor display an image for the resulting workstation to be clinically viable for the radiologists who are used to working with film and alternator? A preliminary experiment [17] suggested that eyetracking of mammographers reading films could yield useful information to help answer these questions.

We conducted an eyetracker study of four experienced mammographers interpreting a variety of cases. An eyetracker is a device that tracks where someone is looking. It allows researchers to determine when the subject is viewing various portions of various images.

2. Materials and Methods

Subjects: Two male and two female board-certified radiologists who are experts in breast imaging and faculty members at our institution served as subjects. As a group, they are responsible for all mammograms read at this institution, as well as the instruction of residents. Subjects ranged in age from 34 to 72.

Equipment. The subjects wore an eyetracker, a device that records eye movements superimposed on a TV signal (NTSC) showing the field of view (Eye Mark Recorder Model V EMR-V NAC, Inc.). The eyetracker system consists of a head-goggle unit and a camera controller unit. The goggle unit is mounted on the head using straps and contains the eyetracking optics and electronics. To record eye movement, an infrared light-emitting diode (950 nm wave-length), which is below the sensory level of the eye, projects a dot of light onto the wearer's cornea. This dot is reflected form the cornea and detected by a video camera (metal-oxide-semiconductor), and finally sent to the camera controller for processing. In addition to the camera for each eye, there is a third "Cyclops" video camera mounted at the center of the forehead that observes the central portion of the subject's field of view. In real time, as the head and eyes move, the camera controller electronically superimposes two eye position indicator spots (e.g., a square) onto the video signal from the Cyclops camera. These spots denotes the instantaneous location of each eye. This combined video signal is available for display on a video monitor and/or recording with a video recorder. The unit has an accuracy of 0.6 degrees with a field of view of 60 degrees horizontal and 45 degrees vertical. The eyetracker output video signal was recorded onto a VHS recorder. In addition, the gross body movements were recorded using a separate camera and recorder. Because the sensory portion of the eyetracker device is mounted completely on the subject's head, subjects are free to move their heads resulting in less interference in the user's behavior.

Cases: To simplify our study, only eight complete cases were used. Each case contained a current and comparison study and each study contained left and right CC and MLO images. These cases were selected to provide a cross section of representative mammographic findings. The cases viewed included: 1. Normal, fatty; 2. Normal, dense; 3. Dominant mass, changing; 4. Dominant mass, stable; 5. Cluster of calcifications, changing; 6. Cluster of calcifications, stable; 7. Multiple bilateral masses; 8. Multiple bilateral calcifications. Three patients in each of the categories were identified using computer records from the years 1993-1994. For the cases

chosen, the patient had to have two consecutive studies done at our institution, separated by at least 12 months; the patient had identifiable mammographic findings; and the films had to be of diagnostic quality. The cases were presented in varying order to each of the four subjects.

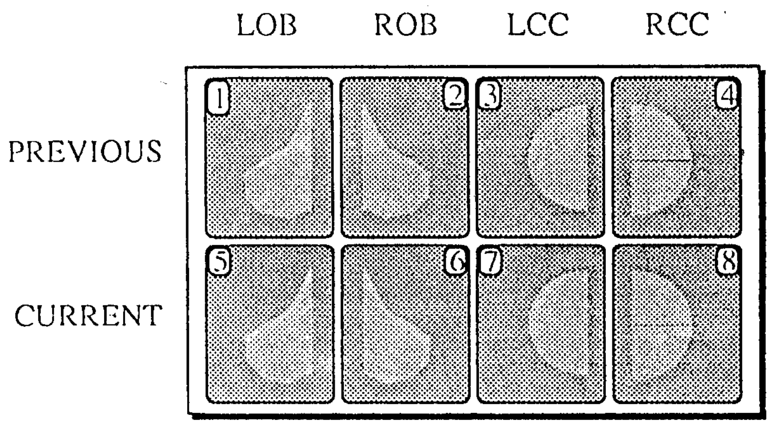

FIGURE 1. Arrangement of mammograms for each study. The numbers are used only for indicating comparison pairs. (L = Left; R = Right; MLO = Mediolateral Oblique view; CC = Cranio-caudal view)

Procedure. In order to provide as realistic an environment as possible, every effort was made to reproduce normal working conditions for the radiologists. The experiment was carried out in the usual clinical setting, the breast imaging reading room at approximately the same time in the afternoon. The most notable difference between our experiment and regular mammogram reading was the presence of the eye tracking device. All films were pre-hung on a RADX dedicated mammography film viewer/alternator. The cases were hung according to the standard practice at our institution (See figure 1). A magnifying glass was available which provided two levels of magnification. Subjects were instructed to generate a clinically acceptable standard mammography report, and were allowed to use the magnifying glass and move images on the board as needed to generate the report. No time limits were imposed. They were given the option to stop the study at any time if they desired. The eye-tracker was calibrated before and after each case using methods supplied by the manufacturer.

Data Collection. The standard mammography report form at the institution was used to record findings. This form provides information to the radiologist on patient demographics (hospital number, age, race), focused history, and current symptoms. It also provided information on

menstrual status and hormonal therapy. The mammographer was required to fill out the section regarding pertinent findings,if there was a significant change noted since the previous study, if the breast parenchyma was dense or fatty, and a list of findings for each breast rated on the ACR 1-5 scale for mammography [ ]. Each of the mammographers was skilled at using this form prior to the study.

Data Analysis. NTSC video generated from the eye-tracker was electronically time-coded with a resolution of 30 frames per second. A video cassette recorder capable of shuttling frame by frame was used to analyze the video (Panasonic SVHS MTS AG-1960) and a high resolution gray-scale monitor was used to view the video. The tape for each trial was analyzed frame-by-frame at a 1/30 second resolution and for each frame, the position of the dominate eye was recorded on paper using a grid pattern indicating the position of all the images in a dual-study case.

The eye tracker device occasionally would slip on the subjects head somewhat during a trial resulting in varying amounts of eye movement inaccuracy for a given trial. This was determined by the calibration sequences performed before and after each case. Thus, for analysis purposes, two levels of eye-movement accuracy were used. *Full-image resolution* noted only which image the eye was viewing in a video frame, while *1/16-image resolution* noted, for a given video frame, not only which image the eye was viewing, but also which segment of a 4x4 grid imposed on that image the eye was viewing. If the post-trial calibration indicated more than a 3 cm variation in eye position from the pre-trial calibration, the trial was deemed to have insufficient accuracy for the *1/16-image resolution* and was thus only used for *full image* resolution. (1/16-image resolution provides a measure of how many roaming operations will be needed to view a 4000x4000 pixel mammogram using a 2000x2000 pixel monitor. The full image resolution, while not providing the 1/16-image roaming information, does provide essential information as to the number, order, and type of image display operations needed to view 8 or more images on 1, 2, 3, or 4 video monitors.)

Workstation users zoom into an image by pressing a button or moving a mouse. In order to be able to predect workstation zoom behaviour, we had to infer from alternator behaviour when the user might zoom with a workstation. Given the roughly 4000x4000 pixel images and 2000x2000 pixel monitors, only a binary zoom would be needed. Thus, the user is either at full resolution, or at 2000x2000 pixel resolution. ROC analysis of digitized film [18] indicates that 2000x2000 pixel images are almost, but not quite, sufficient for mammography interpretation, so mammographers only occasionally need the higher resolution. We thus assumed that the

2000x2000 pixel resolution would be sufficient for all viewing except when the magnification glass was used with film and alternator. When the user is not using the magnification glass, we assumed they were viewing the entire mammogram at 2000x2000 pixel resolution and thus do not need to roam within the image. It is possible that this assumption underestimates the required number of roam and zoom operations. It is also possible that it overestimates that number. Thus our assumptions as to the number of roam and zoom operations are of limited accuracy. Nevertheless, they provide us with a basis for some preliminary conclusions about mammography workstation design.

3. Results

Table 1: Interpretation Times with Eyetracker in Minutes.

Case	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Subject Avg.
Subject A	3.04	1.93	1.73	3.01	2.34	1.12	1.97	3.60	**2.34**
Subject B	4.32	2.37	5.33	2.97	2.42	2.98	2.98	4.47	**3.48**
Subject C	2.54	3.41	1.28	1.48	3.09	2.82	0.87	2.36	**2.23**
Subject D	1.86	1.34	1.23	1.47	1.01	1.00	0.53	1.54	**1.25**
Case Avg.	2.94	2.26	2.39	2.23	2.22	1.98	1.59	2.99	2.32

Data. Table 1 indicates the interpretation times for all the trials in the experiment. All trials were sucessfully completed and allowed for inter-image analysis. However, only 6 cases were analyzed at the 1/16 image resolution: subject B cases 4 and 7, subject C case 2, and subject D cases 1, 2, and 7.

Table 2 contains eyetracker-derived estimates - based on the 1/16 image resolution and full-image resolution - of several workstation operations as the number of workstation monitors varies from one to four. Six cases may seem to be too limited of a sample for a time motion analysis of roam behavior. However, these cases provide for a total of over 80 "roam "data points across three subjects, and are more than sufficient for the simple time motion workstation-design decision-making purposes to which this data might be applied. The information in table 2 was derived from several thousand experimentally gathered datapoints that denoted for each subject and case every 30th of a second, in which 4x4 grid or which image the radiologist was viewing. The three types of operations are as follows:

Average "Image Display" Operations indicates the estimated number of times a particular image would be needed for viewing while not already displayed, thus requiring that image to be called up for display. *Average Image Display Operations* were derived by counting the number of times mammographers moved their eyes from one image to another. As the number of monitors in a workstation is inceased, it is increasingly likely that the desired image to be viewed is already displayed on a monitor. Thus, the number of image display operations decreases as the number of monitors in a workstation is increased from one to four.

Average "Zoom-In" Operations indicates the estimated number of times mammographers either would need to zoom in on an image that was already displayed or display a new image (requiring an image display operation) and zoom into that image. *Average "Zoom In" Operations* were derived by counting the number of times a mammographer picked up a magnifying glass and started looking at an image. Note that we are making an assumption that if the magnifying glass is not being used, mammographers could manage with only 2500x2000 (100 microns/pixel) resolution while they would require a full 4000x5000 (50 microns/pixel) resolution when the magnifying glass was being used. A mammographer may want to zoom into a new image that is already displayed on a monitor and has already been "zoomed". In this case, no zoom operation would be needed. Thus the number of zoom operations decreases as the number of monitors increases.

Average "Roam" Operations indicates the estimated number of times that mammographers would need to move a 2000x2000 pixel viewport on the 4000x4000 pixel mammogram in 1000 pixel increments. We have assumed the 1000 pixel increment as this would allow mammographers to be able to always view any portion of the image with all of its surroundings; a 2000 pixel increment would not allow border pixels to be viewed with pixels just across the border. A mammographer may want to roam to a portion of a new image that is already displayed on another monitor and has already been zoomed and roamed to the required area. In this case, no roam operation would be needed. Thus the number of roam operations decreases as the number of monitors increases.

Table 2: Estimated Workstation Operations for equivalent Interpretation

	1 Monitor	2 Monitors	3 Monitors	4 Monitors
Average "Image Display" Operations	51 (14-92)	29 (12-44)	23 (10-33)	18 (6-30)
Average "Zoom In" Operations	9 (4-17)	7 (4-11)	6 (4-10)	6 (4-10)
Average "Roam" Operations	15 (7-26)	13 (7-21)	13 (7-20)	13 (7-20)

Comparison Pairs. The mammographers often went back and forth between two images presumably looking for differences, similarities, and changes. Table 3 shows the per-case average frequency of viewing for the six most common comparison pairs. A comparison pair is considered to have been viewed when the radiologist views the first image, then views the second, and finally goes back and views the first image. A viewing of an image, followed by a second image and then viewing the first with no intervening viewing of other images would be considered a single viewing of that comparison pair. If the radiologist then went on to view the second image for a second time, that would be considered two viewings of the comparision pair. A third viewing of the first image would be considered a third viewing of the pair.

Only the six most frequently viewed comparison pairs are included in table 3. All other pairs averaged well below 1 viewing per case. As can be seen from tables 2 and 3, display of comparison pairs represents a very significant portion of the total image display operations.

Table 3: Number of Times Comparison Pairs Displayed per Case

			# Times Pair viewed per case
Medial Lateral Oblique	Left	Old & New	3
Medial Lateral Oblique	Right	Old & New	3
Cranio Caudal	Left	Old & New	4
Cranio Caudal	Right	Old & New	5
Medial Lateral Oblique	Left & Right	New	3
Cranio Caudal	Left & Right	New	3

Observations. Although they were given the option, none of our subjects decided to halt the experiment due to discomfort. All of the subjects noted that although the eyetracker device was unwieldy and restrictive at first, it became tolerable and unnoticed as the experiment progressed.

Only one subject complained of any side effects, namely a headache that went away soon after the experiment. Nevertheless, it is possible that the device effected mammographer behavior.

4. Discussion

Number of monitors. Table 2 clearly indicates that increasing the number of monitors will allow a decrease in the total duration of the interpretation. To illustrate this with an example, suppose an *image display* operation (including both hand-motions and system response time) requires 3 seconds and a *zoom* operation or roam operation requires 1 second. From table 1, we can determine that a 1 monitor system would have an average of 51*3+9*1+15*1= 177 seconds or about 3 minutes of *image manipulation time*. Using these same operation durations, the two monitor system would have 29*3+7*1+13*1=107 seconds or about 1.8 minutes of image manipulation time for a 60% reduction over the one monitor system. Moving to a four monitor system would require 18*3+6*1+13*1=73 seconds or about 1.2 minutes of image manipulation for a 40% reduction from the two monitor system. Note that even with reduced duration of the various operations, more monitors will result in a faster interpretation, though the advantage is less with faster operations.

Obviously four monitors would greatly increase the expense of a mammography workstation and also the amount of space occupied in the clinic. Further, modern 2000x2000 pixel monitors are large, and viewing and comparing images on four monitors might require mammographers to move their chairs back and forth between the monitors, increasing the duration of the interpretation in ways not accounted for in the above analysis. Ideally, manufacturers would produce smaller monitors tailored to mammography and package them to reduce the distance between active screen areas.

System Response Time. Image display operations, zoom operations, and roam operations all require the mammography workstation to move a portion of a mammogram onto a particular monitor from a framebuffer, from the workstation's fast random access memory, or from disk. System response time for image display can range from 0.1 to 2 or even 5 seconds with many current medical image workstations. To take an example, suppose a two monitor system has 49 operations (29 image display , 7 zoom , and 13 roam) then a 5 second system response time would result in a 245 second overhead, a 2 second system response time would result in a 98 second overhead, and a 0.1 second system response time would result in a 4.9 second overhead. Clearly system response times of a few tenths of seconds are essential if we are to construct a mammography workstation that can compete with a lightbox.

Table 4: The "interpretation overhead" results of various system response times and monitor configurations in seconds.

Response time	1 Monitor(75)	2 Monitors(49)	3 Monitors(42)	4 Monitors(37)
5 seconds	375	245	210	185
4 seconds	300	196	168	148
3 seconds	225	147	126	111
2 seconds	150	98	84	74
1 seconds	75	49	42	37
0.5 seconds	37	4.1	3.5	3.1
0.1 seconds	7.5	4.9	4.2	3.7

Note: The figure in parentheses shows the number of image operations (image display, zoom, roam) from Table 2.

Table 4 shows the benefits as the number of monitors is increased and as the system response time is decreased. Two monitors with a 0.1 second response time are much faster than a four monitor system with a 2 second response time. Note that response time is only a portion of the overhead for a image display, zoom, or roam operation. The time for the mammographer to move a mouse or press a button can be very significant, and would likely add from 0.1 to 2 seconds to each operation and thus would tend to increase the importance of having a larger number of monitors with the corresponding fewer number of interaction operations.

Comparison Pairs. Displaying a mammogram on a particular monitor normally requires the mammographer to select the image, select the destination monitor, and wait for the system to display the image; these three steps are ergonomically complex and can easily take 3 to 5 seconds for one image and from 6 to 10 seconds for a pair of images depending on required hand motions and system image display time. However, display of a comparison pair takes considerably less time not only because one operation will display both images, but also because (presumably) the workstation designer can a priori determine which image should go on which monitors for comparison of a particular pair of images, eliminating the need for the radiologist to select monitors every time the pair is to be displayed. Thus we roughly estimate that display of a comparison pair can take from 0.5 to 2.5 seconds depending on the workstation's image display

time. Table 3 indicates that considerable ergonomic savings can be achieved by a mammography workstation providing *one-button* function for display of each of the listed comparison pairs. Note that the comparison pair data does not account for all the image display operations, so a conventional mechanism for displaying a particular image on a particular monitor will still be required. The cost of a comparison-pair display function is the increase in complexity and thus the learning time for a mammography workstation.

Workstation Viability. Can we construct a viable mammography workstation using 2000x2000 pixel monitors to interpret eight 4000x4000 pixel mammograms? A reasonable initial goal would be to have the difference between the workstation interpretation time and prehung film/alternator time to be no more than the average time to load the images onto the alternator and to return them back into the folder, say 20 seconds or so. Table 3 indicates that with a 0.1 second image display time and minimum of two monitors we can reduce the time for the computer to display the various images onto the monitors to less than 5 seconds of the 20 second limit. If the handmotions to initiate a roam, zoom, or comparison-pair operation were limited to two button presses or about 0.4 seconds, for a total of 0.5 seconds per operation including the 0.1 second image display time, the total workstation overhead for a four monitor workstation with its estimated 37 operations (table 4) would be 19 seconds, which might just produce a viable mammographic interpretation environment given the improved logistics of the filmless environment.

Caveats. There are a number of circumstances that somewhat limit the applicability of this study. First, wearing the eyetracker device and knowing they were being observed almost certainly affected the behavior and speed of the mammographers. Second, only 6 of the 32 trials were analyzed at the 1/16 image level of detail, though we believe that the number of data points analyzed were sufficient to make our limited inferences. (Zoom and image display operations were derived from all 32 trials.) There were eight images each on those 6 trials and together these represent over 80 roaming operations. Further, these 6 trials represented varying subjects and cases. It is possible that an increase in the number of trials analyzed at the 1/16 image resolution would have resulted in somewhat different numbers. However, given the inherent inaccuracies and limitations of time motion analysis to which these numbers will be applied, the 6 trials should be more than sufficient for comparison of various "roam" design alternatives. Third, as mentioned in the introduction, we have ignored the effect of greyscale manipulations on the ergonomics of workstation in general and on its viability for mammography in particular. If workstation display of digital mammography requires intensity windowing while film display does not, then film will have a significant advantage as the number of images to be viewed by the

mammographer may be doubled or even tripled by workstation display, although a possible improvement in interpretation quality with grey scale manipulation available through the use of a workstation might also justify any increased interpretation time.

References

1. Shapiro S, Venet W, Strax P, Venet L, Roeser R. Ten to fourteen year effect of screening on breast cancer mortality. *JNCI* **1982**; 69:349-355.
2. Tabar L, Fagerberg G, Duffy SW, Day NE. The Swedish two county trial of mammographic screening for breast cancer: reent results and calculation of benefit. *J Epid Com Health* **1989**; 43:107-114.
3. Shapiro, S, Venet W, Strax P, Venet L (eds.). *Periodic screeing for breast cancer*. Baltimore, Johns Hopkins, **1988**.
4. Baker LH. Breast cancer detection demonstration project: five-year summary report. *CA* **1982**; 32(4): 194-225.
5. Saltzstein SL. Potential limits of physical examination and breast self-examination in detecting small cancers of the breast: an unselected opoulation-based study of 1302 cases. *Cancer* **1984**; 54: 1443.
6. Fletcher SW, O'Malley MS, Bunce LA. Physicians abilities to detect lumps in silicone breast models. *JAMA* **1984**; 251, 1580.
7. Fisher B, et al. Cancer of the breast: size of neoplasm and prognosis. *Cancer* **1969**; 24:1071.
8. Wanebo HJ, Huvos AG, Urban JA. Treatment of minimal breast cancer. *Cancer* **1974**; 33:349.
9. Gallager HS, Martin JE. An orientation to the concept of minimal breast cancer. *Cancer* **1971**; 28:1505.
10. Frazier TG, Copeland EM, Gallager HS et al. Prognosis and treatment in minimal breast cancer. *Am. J. Surg* **1977**; 133:697.
11. Kimme-Smith C, Bassett LW, Gold RH, Zheutlin J, Grombein JA. New Mammography Screen/Film Combinations: Imaging Characteristics and Radiation Dose. *AJR* **1990**;154:713-719.
12. Haus AG. Technologic Improvements in Screen-Film Mammography. *Radiology* **1990**; 174(3):628-637.

13. Stomper PC, Gelman RS. Mammography in symptomatic and asymptomatic patients. *Hem/Onc Clinics NA*, **1989**; 3(4):611-640.

14. ACR manual 1993

15. Shtern, F. Digital mammography and related technologies: a perspective from the National Cancer Institute. *Radiology* **1992**; 183:629-630.

16. Maidment ADA, Yaffe MJ. Scanned-slot Digital Mammography. *SPIE*, **1990**; 1231: 316-326.

17. Beard DV, Johnston RE, Pisano E, Hemminger B, and Pizer SM. An Electronic Lightbox for Mammography: Preliminary Observations, Eyetracker Studies, and Design. *SPIE Medical Imaging V* 1991; 1446: 289-296.

18. Oestmann JW, Kopans D, Hall DA, McCarthy KA, Rubens JR, and Greene R. A Comparison of Digitized Storage Phosphors and Conventional Mammography in the Detection of Malignant Microcalcifications. *Invest Radiol* **1988**; 23: 725-728.

List of figure and table captions:

FIGURE 1. Arrangement of mammograms for each study. The numbers are used only for indicating comparison pairs. (L = Left; R = Right; MLO = Mediolateral Oblique view; CC = Cranio-caudal view)

Table 1: Interpretation Times with Eyetracker in Minutes.

Table 2: Estimated Workstation Operations for equivalent Interpretation

Table 3: Number of Times Comparison Pairs Displayed per Case

Table 4: The "interpretation overhead" results of various system response times and monitor configurations in seconds.

APPENDIX B

Does CLAHE Improve Detection of Spiculations in Digitized Mammograms?

William F. Garrett[†], Etta D. Pisano[*], Derek T. Puff[°]

From the [†]Department of Computer Science, [*]Department of Radiology, and [°]Department of Biomedical Engineering

University of North Carolina-Chapel Hill, Chapel Hill, NC, 27599-7510

INTRODUCTION

Screening mammography has proven to be an effective test in identifying early breast cancer. Randomized trials have demonstrated that, for women over age 50, breast cancer mortality can be reduced as much as 30% through mammography and breast physical examination [8]. Unfortunately, as many as 10% of palpable breast cancers are not visible with standard mammographic techniques. Our aim is to improve the accuracy of mammography with digital image processing.

We conducted two laboratory experiments to determine the potential benefit of Contrast Limited Adaptive Histogram Equalization (CLAHE) to mammography. Our goals were twofold: first, to determine if CLAHE could improve the detection rate of simulated spiculations in mammograms; second, to determine the choices of CLAHE parameters that yield the best enhancement. This paper describes our methods, results, and conclusions.

I. BACKGROUND AND SIGNIFICANCE

A mammogram is generated by shooting an x-ray photon beam through the patient's body and onto a film-screen system. The x-rays are attenuated by the bodily tissue they pass through before they strike the screen. Photons striking the screen cause it to emit visible light that exposes the film. Dense tissue attenuates the beam, resulting in a lighter (brighter) image on film. If no photons are stopped, the film appears black.

A mammogram consists of four images: left and right craniocodal views (taken from above the head looking down), and left and mediolateral oblique views (side views). Radiologists use these four views, as well as sets of images from previous examinations of the same patient, in their analysis.

I.A What radiologists look for in an image

In general, radiologists look for five features in a mammogram as possible indications of cancer: *masses*, *spiculations*, *calcifications*, *architectural distortions*, and *asymmetries*. *Masses* typically looks like rounded or oval lumps 6 - 10 mm in diameter, with curved borders. They may appear brighter than surrounding tissue because they are more dense, or they may be the same intensity. They can have sharp or ill-defined edges, with less well-defined masses being more characteristic of malignancy. *Spiculations* are small tendrils that grow from cancerous tumors. Radiologists look for spiculations to determine whether a tumor is malignant or benign. Sometimes the spiculations are more visible than the mass; they are often spotted because they don't necessarily run in the same (center-ward) direction as the rest of the breast tissue. *Calcifications* (also known as microcalcifications) are small calcium deposits that typically appear in small clusters. Their presence may indicate breast cancer. The fourth type of feature is an *architectural distortion*. While a mass may not be visible in an image, the presence of a mass in the breast can displace some of the surrounding tissue from its usual gracile arcs extending toward the nipple. These distortions can also be spotted *asymmetries* between the left and right breasts.

I.B. Previous image enhancement work in radiology

Few investigators have studied the application of digital image processing techniques to mammography. McSweeney tried to enhance the visibility of calcifications by using edge detection for small objects, but never reported any clinical results [7]. Smathers showed that intensity band-filtering could increase the visibility of small objects compared to images without such filtering [12]. Chan used unsharp masking (an edge-sharpening technique used in photography for many years) to remove image noise for computerized detection of calcification clusters [1]. Chan noted that while these techniques improved detection, the improvements may have been greater if the observers had been trained to make diagnoses from the processed mammograms rather than the unprocessed (normal) mammograms [2].

Previous work at UNC has explored the use of Intensity Windowing (IW) and the Adaptive Histogram Equalization (AHE) family of algorithms in mammography and computed tomography [6,9,10]. Puff described a method for using CLAHE to improve the detection of masses [11]. An important conclusion of his study was that radiologists and non-radiologists exhibit similar trends in detection performance. While non-radiologists did not perform as well as radiologists overall, the two populations displayed parallel increases and decreases in performance due to image processing. We use this in our study as a justification for selecting non-radiologist observers.

Puff's work, combined with pilot studies in our group and the results of other research groups, suggests that different image processing methods may be better suited for the enhancement of certain features than others. We believe,

from both pilot analysis and mathematical understanding of the algorithm, that CLAHE may be most applicable to the enhancement of spiculations.

I.C. How CLAHE works

CLAHE is a member of the AHE family of algorithms developed at UNC by Stephen Pizer in the early 1980s [9]. It is an adaptive contrast enhancement technique that alters image pixel intensities as a function of the intensities of neighboring pixels. CLAHE has two parameters: *region size* and *clip limit*. Region size is the size of the neighborhood of pixels that are used in the recalculation of one pixel's intensity. Clip limit restricts the amount by which CLAHE can alter the intensities of the image; the desirability for this limiting is described below. We present a brief description of how CLAHE works. Readers who desire a more detailed explanation should consult [4,5,9]. Our description loosely follows that of [4].

I.C.1. Global contrast enhancement

A global or stationary enhancement mapping is a gray-level transformation in which the intensity of each pixel in a digital image is altered according to a mathematical function of intensity values. The goal of the algorithm designer is to find a function that best utilizes the full range of displayable gray levels. Intensity windowing (IW) and histogram equalization are examples of global mapping.

With intensity windowing, we define a subrange of gray levels and expand it linearly to fill the full range of available intensities. This dramatically enhances the contrast of features falling within the specified window of input intensities, but maps pixels outside of that range to the minimum or maximum levels. IW is one of the fastest global algorithms; it can usually be computed in the display subsystem by manipulating the lookup tables of the display device. Although the parameters of IW can be adjusted interactively, preset values are often used. In chest CT, for example, radiologists are often presented with a "lung window" preset, a "mediastinum window" preset, etc.

Global histogram equalization maximally transmits the visual information contained in image intensity values [4]. The algorithm constructs a histogram of intensity levels and computes a mapping in which a pixel's new intensity is proportional to its rank in the histogram of the entire image. The mapping is thus proportional to the cumulative distribution function of the image intensities. That is, the intensity values represented in the greatest number of pixels in the input image are mapped to the greatest number of display levels in the output image.

In a more intuitive sense, the object of interest in an image often covers a sizable portion of the image but is represented by only a narrow range of gray levels. For example, a lung in a chest CT scan or breast tissue in a mammographic image, will be displayed in a narrow range of intensities because the tissue is fairly uniform in density. Histogram equalization will assign a greater range of intensity levels to these large objects, making them much more visible. The "information" about these features was always present in the unprocessed image but was effectively hidden from the human eye because it lacked perceptible

contrast. Figures 1a-b show an unprocessed CT scan and the same scan processed with histogram equalization.

A B C

FIGURE 1: A. Plain film chest CT scan. **B.** Same scan, processed with histogram equalization. Notice how more of the structure of the internal organs is visible. **C.** Same scan, processed with CLAHE. Note the greatly improved visibility of internal structures. Chest CT images are shown rather than mammographic images because mammograms are too big to include in this manuscript.

Global contrast enhancement has three shortcomings. First, portions of the image that were discernible before processing are mapped to uniform black or white as part of the processing. Information in those portions is lost. Second, parts of the input image occupying widely separated areas of the intensity range cannot be enhanced effectively in the same output image. Third, and perhaps most seriously, the perception of object boundaries can depend critically upon window selection [4]. Adaptive contrast enhancement techniques address these concerns.

I.C.2. Adaptive contrast enhancement

Adaptive contrast enhancement algorithms map intensity values based on their original values (as with global enhancement) and local image characteristics in a certain contextual region. In AHE, this region is a square of some number of

pixels in width, centered upon the pixel being transformed. A new region is considered for each pixel in the input image. In our experiments we used an approximate CLAHE algorithm, one that uses static regions of fixed size and location. With this method, a window of n by n pixels is used in the remapping of all n^2 pixels in that region. Although the results of the processing are not the same as with the true CLAHE algorithm, the approximation runs significantly faster and produces nearly the same results.

In AHE, an intensity histogram is calculated for each region of the image, and the image transformation equalizes this local histogram. This approach is logical both from the point of view of information theory and from our knowledge of the human visual system: humans are very sensitive to local relative contrasts but insensitive to both absolute luminance and the contrast of images separated by a large physical distance.

One problem with AHE is that it has no concept of signal or noise; noise is enhanced along with the rest of the image. In addition, it can over-emphasize strong edges, making it hard to determine where the edges are really located. CLAHE tries to solve these problems by limiting the intensity map calculation according a user-specified clip limit parameter. This value controls the maximum height of the intensity histograms calculated in the algorithm; where the histogram exceeds the maximum value, it is clipped to the maximum value. The lower the clip limit, the less the effect of the remapping function. This helps prevent the over-enhancement of noise and reduces the edge overshoots of unlimited AHE. Figures 1a-c show how CLAHE improves the visual quality of images over unprocessed images and images processed with global histogram equalization.

I.D. How we intend CLAHE to be used and what we expect

We expect CLAHE to improve mammographers' diagnostic accuracy. Our experiment was performed in a lab rather than a real clinical setting because we wanted to control the statistical power of the experiment and better control the variables, all within a reasonable period of time.

If and when CLAHE is used in the clinic, we intend it to be used as an adjunct method, not a replacement for standard mammographic images. There are two reasons for this. First, images processed with CLAHE differ greatly in appearance from the standard images that radiologists are accustomed to seeing. Second, CLAHE enhances noise and can produce images that are worse (for mammographic analysis) than the originals. We do not attempt to compensate for this; we readily admit that some combinations of parameters will produce images that are worse than the originals. The goal of our experiment is to identify the settings that produce significantly *better* images in a wide variety of cases.

II. MATERIALS AND METHODS

Our study required observers to determine the orientation of a simulated feature embedded in a real image. Their accuracy of detection over different cases of image processing was used to determine the improvement offered by CLAHE. We conducted two experiments that differed only in parameter choices for the stimuli. This section of the paper describes how we generated the stimuli, what

the observers' task was, how we conducted the experiment, and how we analyzed the data we collected.

II.A. Preparing the stimuli

We wrote a computer program to construct the stimulus images. It randomly selected one of four background images and rotated that background to one of four orientations. These four images and four orientations provided 16 essentially different backgrounds. Next, the program added a phantom feature (a spiculation) into the background. The image was processed with CLAHE to yield the final stimulus. The program grouped stimuli into 20 grids of 8x8 images each and stored them on disk. The grids were printed onto film for use in the experiment. These steps are detailed in the following sections.

II.A.1. Selecting images

We used four background images of 256x256 pixels each, cropped from actual clinical mammograms digitized with a 50 micron sample size with 12 bits (4096 values) of intensity data per sample. The images contained relatively dense breast parenchyma. They were known to be normal (no evidence of cancer) by previous examination. Figure 2a shows one of the backgrounds. We used the same set of images for both experiments.

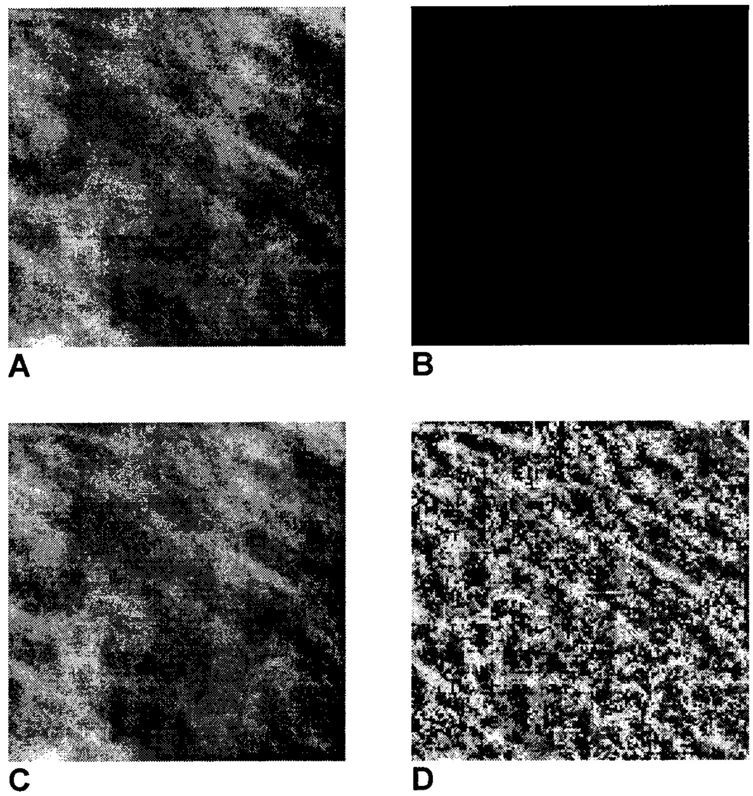

FIGURE 2: A. A 256x256 region cropped from a digitized clinical mammogram. **B**. A simulated spiculation. **C**. The image from A with the feature from B added in by pixelwise addition. **D**. The same image, processed with CLAHE. Notice how the feature is more visible after processing.

II.A.2. Modeling the features

After selecting and rotating the background for a stimulus, the program inserted a phantom spiculation. We simulated mammographic spiculations as 1 pixel-wide lines approximately 11 mm. in length. They were positioned at orientations of 0, 45, 90, and 135 degrees, passing through the center of the stimulus. Positions were jittered randomly by a few pixels so that the stimuli would not always appear in the same four places. The features were embedded in the images by pixelwise addition with the backgrounds. Figure 2b shows an example of a simulated spiculation; figure 2c shows the background image with the spiculation added in. We used simulated features instead of real features so

that we could have precise control over the location, orientation, and figure-ground contrast of the spiculations.

The features were added at specific contrast levels (10, 25, 40, and 55), given in terms of the number of gray levels between background and foreground. Normally, contrast is a dimensionless measure, formed by a ratio of foreground and background luminances or gray levels. Because we performed pixelwise addition, however, and because the intensity of the background varies due to structure and noise, such a ratio is difficult to formulate. Thus we use the gray-level measure instead.

Moreover, our definition of contrast creates an independent variable with discrete levels for analyzing perceptual thresholds. We performed pilot studies to choose contrast levels that would best characterize the visual response curve (described later) and used the same set of contrast levels in both experiments.

II.A.3. Processing the images

We used an approximation CLAHE algorithm to process images because it is computationally faster than the true CLAHE algorithm. The approximation CLAHE routine takes parameters for clipping level and number of regions, where the number of regions is equal along each axis and must be square. Real CLAHE takes region size instead of number of regions. Number of regions approximately corresponds to image size divided by region size.

Both experiments used 10 combinations of CLAHE parameters. One was the control (no processing); the remaining nine were combinations of three clip limits

and three region sizes. In the first experiment, we used clip limits of 8, 16, and 32; and region sizes of 16, 32, and 64 pixels in each dimension. We chose these CLAHE parameters after analyzing our pilot studies, selecting values that spanned the effective range of the parameters and represented the best enhancement choices. In the second experiment, we used clip limits of 2, 4, and 16; and region sizes of 8, 32, and 128. The first set of data had indicated a positive trend in detection with lower clip limits and larger regions; we selected the second set of parameters to better explore these trends. Figure 2d shows an example of a stimulus image processed with CLAHE.

Each experiment comprised 1280 stimuli selected from the 2560 possible combinations of all parameters. Of these, the 40 combinations of contrast level (4 contrasts) and processing condition (10: 9 plus no processing) were the key parameters. We generated 32 unique trials for each of these combinations by selecting 32 of the 64 combinations of background image (4 images), background rotation (4 angles), and feature orientation (4 angles). Images were organized randomly into the grids.

II.A.4. Printing images onto film

The digital images obtained from the computer program were printed onto standard 11x14 mammography film. We calibrated the printer so that its input driving levels corresponded roughly linearly to optical density on the resulting films. This transfer function is nonlinear only at the high and low extremes. It is important that this function be as linear as possible so that the contrast of the features in the printed images is the same as the contrast in the digital images we generated on the computer.

We presented the images on film rather than on a computer monitor for several reasons. First, monitors are incapable of displaying the intensities lightboxes can display. A typical monitor achieves 5 foot-lamberts of illumination, while a standard lightbox can achieve several hundred. Moreover, monitor display requires adjustment for intensity linearization. Second, monitors do not have the same spatial resolution as film. A typical workstation monitor displays 100 pixels/inch, whereas the film provides approximately 300 samples/inch. Until monitors can match lightboxes in intensity and resolution, radiologists will use lightboxes. This is the third reason we chose to use lightboxes: they will be the standard tool in mammography for several years to come.

II.B. Experimental procedure

II.B.1. Observers

Our investigations were conducted with an observer population consisting primarily of graduate students from the medical school, biomedical engineering department, and computer science department. We sought people who had some familiarity with medical imaging, but specifically excluded those too familiar with mammography. While these naïve observers are not as accurate in detecting the features as experienced radiologists, our previous work shows that both groups demonstrate the same trends in detection across contrast and type of image processing [11]. We used the student observers because we needed two dozen subjects for approximately 5 hours each. Trained radiologists are not so readily available.

Observers were paid for participating in the study. They received a flat amount for completing the experiment plus a variable bonus keyed to their accuracy in answering. We intended the performance bonus to motivate the observers to answer as accurately as possible rather than finish as quickly as possible. All observers signed an informed consent form after we explained to them the nature of the study. There were 10 observers in the first study and 13 in the second.

II.B.2. The task

Observers had to view each image to determine the orientation of the phantom feature in each. They chose from answers depicting the four orientations used in the experiment (a 4-AFC paradigm). Observers were instructed to make their best guess if they were unsure of the orientation of the feature.

Films were displayed on a standard mammography lightbox that was masked with heavy paper so that only the grid of images on the film was illuminated. The experiment was conducted in a visual perception laboratory with controlled lighting. Observers used a standard mammography magnifying glass to view the images.

We trained each observer for the task in the experiment. The training session comprised a brief explanation of the purpose of the study, a description of what each stimulus represented, instructions for performing the experiment, and two training sets of images which the observer analyzed with immediate feedback from the experimenter.

The first training set was "easy", having structures of high contrast with the background. It was intended to familiarize observers with the task. The second contained structures of the same contrast levels as the actual experiment, and was designed to acclimate observers to the experiment's difficulty. The experimenter provided immediate feedback on these practice sets, telling observers the correct answers and helping them learn how to spot the features. The experimenter left the observer alone after the training session, returning periodically to monitor his/her progress.

Observers were instructed to take breaks as often as necessary, and at least once every half hour. Because the experiment was conducted on film (as opposed to computer monitor), and because we demanded a positive answer to each trial, we could not enforce a viewing duration for each stimulus. Observers were instructed, however, to spend approximately 5 seconds on each stimulus, regardless of its difficulty. The experimenters monitored the observers' progress with periodic checks and encouraged them to maintain that pace. Overall, the experiment took 4-5 hours for each observer, divided into two sessions of approximately 150 and 120 minutes.

II.B.3. Collecting Data

Observers marked their answers by hand on a paper answer sheet that contained a grid of the same size and shape as the grid of images on film. They drew lines inside the grid boxes to indicate the orientation they believed each stimulus to have. The experimenter collected these answer sheets and entered them into the computer.

The program that created the images also created a database of the values of each variable in each image. The data-entry program accessed these record files, so that the experimenter only needed to enter each observer's answers, and the program automatically associated each answer with a full listing of the parameters in the corresponding stimulus. These records were stored on disk for later statistical analysis.

II.C. Statistical analysis

We randomly varied feature contrast and the two CLAHE parameters in order to derive a relationship between these three parameters and accuracy of image detection. Different sets of image processing parameters can be compared by evaluating the shifts they cause in the curve relating contrast to accuracy of perception for each set of CLAHE parameters.

II.C.1. Contrast perception and psychometrics

Our statistical analysis relates perception of a feature to the perceptual contrast between that feature and the surrounding background. The contrast levels we dealt with in producing the stimuli were not perceptual contrasts; rather, they were differences in digital driving levels. These driving levels map to optical film opacity in the printing process, and film opacity maps to optical intensity when the films are displayed on a lightbox. Fortunately, both of these processes are essentially linear: the transformation from driving levels to film in the printer, and the transformation from opacity to intensity with the lightbox. Physics guarantees the latter, and we calibrated the printer to guarantee the former. Thus, we

consider the transformation from driving levels to intensities to be linear. We model perceptual contrast as the logarithm of intensity and use that log quantity for our statistical analysis. This assumption is widely accepted in the field of Human Vision [3].

Classical sensory discrimination theory predicts that, since contrast values were varied from virtually imperceptible to highly apparent, a typical S-shaped curve will describe the data [3]. Detection performance for features well below the perceptual threshold asymptotically approaches 25 percent in a 4-alternative forced choice (4-AFC) paradigm because observers can, by chance alone, guess the answer correctly in 1 out of every 4 trials. Likewise, performance on high-contrast features asymptotically approaches 100 percent because the features are readily apparent.

II.C.2. Probit analysis

We analyze perceptual response with a probit model, a method that models a proportion outcome (percent correct, in this case) as a function of a continuous predictor (in this case, feature contrast). Probit analysis assumes a cumulative Gaussian (normal) distribution model, yielding values for the mean and standard deviation parameters that describe the Gaussian distribution.

The mean parameter, μ, indicates the inflection point of the sigmoidal probit curve. This parameter is counted in digital gray levels. As its value increases, performance accuracy decreases because the detection curve is shifted to the right, meaning that higher contrasts are required for the feature to be visible.

Large values for the standard deviation parameter, σ, indicate a small (shallow) slope of the function.

We modeled the data with a different value of μ for each observer and processing condition, but used a single value of σ per observer. Previous investigations [11] have shown that a stable numerical solution to our statistical analysis is not possible when we attempt to fit a distinct value of σ to each subject and processing condition. We analyzed the logarithms of the parameters; that is, we computed statistics based on $\log_2$(clip size) and $\log_2$(number of regions).

For our analysis we defined two candidate measures. First, a response variable "umstd" that measures standardized inverse mean:

$$umstd_{i,j} = (2-\mu_{i,j})/\sigma_i$$

In this expression, i denotes the subject, j denotes the processing condition, and μ and σ are the probit curve parameters. Subtraction of μ from 2 inverts the function, yielding the intuitive schema of a larger score representing greater accuracy in detection.

We wanted a measure to better describe μ and σ, though, so we devised a second statistical measure, the theta score:

$$\theta = \mu_{i,j} + \sigma_i$$

Because we are interested in the improvement offered by CLAHE, we measure the "success" of an processing condition by the difference between its theta score and the theta score for the unprocessed case. A large theta score indicates that a filtering worsened performance because the observer could only detect the easier (higher contrast) stimuli. A large difference-of-theta score reflects improved performance because it indicates better detection with processed images than with unprocessed images.

Repeated measures Analysis of Variance (ANOVA) was performed on the umstd and theta scores to test differences between processing conditions and observers. In addition to describing the effects of manipulating single variables, ANOVA describes interactions between variables.

III. RESULTS

We began with a univariate approach to repeated measures. Our first test was to determine if there was an interaction between log(clip) and log(regions). Geisser-Greenhouse analysis showed that there is an interaction between the two variables ($p=.0026$, G-G $\varepsilon = .7204$). Next we performed a series of step-down tests to determine the nature of the interaction. We tested four candidate interactions: quadratic in log(clip) by quadratic in log(regions), quadratic in log(clip) by linear in log(regions), linear in log(clip) by quadratic in log(regions), and linear in both variables. Table 1 shows the significance of each of these hypotheses.

TABLE 1: Interaction Between Variables

Candidate measure	F Value	Pr > F
$[\log_2 Clip]^2 \times [\log_2 Regions]^2$	3.39	0.0904
$[\log_2 Clip]^2 \times \log_2 Regions$	15.70	0.0019*
$\log_2 Clip \times [\log_2 Regions]^2$	6.32	0.0272*
$\log_2 Clip \times \log_2 Regions$	2.07	0.1760

We allowed an error of 0.04 on this test, so the two quadratic-by-linear interactions (marked with asterisks in Table 1) were accepted.

We ran a second series of tests to determine if there is a significant difference between the scores from processed images and the scores from unprocessed images. We allowed 0.01 error on this test, so that the total error between this test and the previous test is 0.05. This test has nine separate hypotheses, though, corresponding to the question of whether each of the nine processing cases offers an improvement over unprocessed images. We used a Bonferroni correction to control the overall error rate, giving us an allowable error of 0.0011 on each individual case. Table 2 shows the results of a T-test to determine the validity of each hypothesis.

TABLE 2: Difference Between Processed and Unprocessed Cases

log(Clip)	log(Regions)	Mean	Std. Dev.	T	Prob > |T|
1	1	0.12	0.075	5.59	0.0001*
1	3	0.17	0.074	8.36	0.0001*
1	5	0.17	0.088	6.77	0.0001*
2	1	0.08	0.073	4.05	0.0016
2	3	0.17	0.079	7.71	0.0001*
2	5	0.19	0.082	8.38	0.0001*
4	1	0.13	0.076	6.35	0.0001*
4	3	0.15	0.100	5.66	0.0001*
4	5	0.22	0.078	10.01	0.0001*

The final column lists the probability of this hypothesis being correct for each processing case. Asterisks show which meet our self-defined criteria for significance.

As Table 2 clearly demonstrates, we can refute the hypothesis that there is no difference between processed and unprocessed images for eight out of the nine cases (marked with asterisks in Table 2). That is, we can conclude that those cases offer a significant overall improvement over unprocessed images.

Which processing conditions are best? Table 3 shows the theta scores for the processing used in both experiments.

TABLE 3: Difference of Theta Scores

	log(Clip)				
log(Regions)	1	2	3	4	5
1	.12[b]	.08[b]		.13[b]	
3	.17[b]	.17[b]		.15[b]	
4			.15[a]	.19[a]	.15[a]
5	.17[b]	.19[b]	.19[a]	19[a], 22[b]	.20[a]
6			.18[a]	.14[a]	.17[a]

Scores marked with (a) are from the first experiment, (b) are from the second experiment.

IV. DISCUSSION

These results are encouraging. The probit model predicts that in a clinical environment (i.e., ∞-AFC) CLAHE processing will increase detection rates by as much as 35% in cases near the threshold of detection. This is the first experiment in mammography (in the authors' knowledge) that demonstrates that an algorithm improves the accuracy of detection in a laboratory setting. We hope that CLAHE will improve detection in the clinic, and hence lead to more accurate diagnosis of breast cancer patients. The results suggest increased sensitivity of spiculations, i.e., better detection through fewer false negatives.

Despite its promising results, however, this study has several limitations. First and most importantly, it is a lab study rather than a clinical study. Clinical studies will have to be performed before CLAHE processing is made a routine procedure. Second, this study was not conducted with radiologists as subjects. While we have found that radiologists and graduate student observers demonstrate the same trends in accuracy, it might be the case that radiologists

are already doing so well in practice that CLAHE will not significantly improve their performance. Future experiments will need to involve real radiologists. Finally, our simulated features may be inaccurate. We continue to explore methods of simulating features, but the best solution to the question of accuracy is to use real features. We have avoided them in the past because they cannot be manipulated like simulated features.

Clearly, a clinical study is needed, although it will take much longer than our lab studies. A prospective study would apply CLAHE to real cases and would measure detection rates over many cases and many doctors. Such a study would take as many as 5 years and would probably have to be a multi-center effort. Breast cancer simply isn't a common disease. Even among screening patients, the rate is only 7 per 1000 women at their first screen, and 4 per 1000 women at subsequent screens. Moreover, the cases that might be benefited if CLAHE improves detection as suggested in this study — spiculation cases that were missed in normal analysis — form only a small subset of this set of all cases.

Finally, by its design, this study did not assess the impact of this algorithm on the specificity of mammography. It might be that the addition of CLAHE to the mammographer's tools could significantly increase false positive examinations and do more harm than good. Clinical trials are necessary before we can assess the impact of this image processing algorithm.

REFERENCES

1. Chan HP, Doi K, Galhorta S, Vyborny CJ, MacMahon H, Jokich PM. Image feature analysis and computer-aided diagnosis in digital radiography: I. automated detection of microcalcifications in mammography. *Med. Phys.* **1987**; 14, 4: 538-547.

2. Chan HP, Vyborny CJ, MacMahon H, Metz CE, Doi K, Sickles EA. Digital mammography ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications. *Investigative Radiology* **1987**; 22: 581-589.

3. Corso JF. *The experimental psychology of sensory behavior.* New York: Holt, Rinehart, and Winston, 1967.

4. Cromartie R, Pizer SM. Structure-sensitive adaptive contrast enhancement methods and their evaluation. *Image and Vision Computing* **1993**; 11, 8: 460-467.

5. Cromartie R, Pizer SM. Edge-affected context for adaptive contrast enhancement. *Proceedings of the 12th International Conference on Information Processing in Medical Imaging* **1991**; 474-485.

6. Hemminger BM, Johnston RE, Muller KE, Taylor D, Mauro M, Schiebler M. Comparison of clinical findings between intensity-windowed versus CLAHE presentation of chest CT images. *University of North Carolina at Chapel Hill, Technical Report* TR93-024 **1993**.

7. McSweeney MB, Sprawls P, Egan RL. Enhanced image mammography. *American Journal of Radiology* **1983**; 140: 9-14.

8. Pisano ED, Shtern F. Image processing and computer aided diagnosis in digital mammography: a clinical perspective. *Intl. J. of Pattern Recognition and Artificial Intelligence* **1993**; 7, 6: 1493-1503.

9. Pizer SM, Zimmerman JB, Staab EV. Adaptive grey level assignment in CT scan display. *Journal of Computer Assisted Tomography* **1984**; 8, 2: 300-305.

10. Puff DT, Cromartie R, Pisano ED. Evaluation and optimization of contrast enhancement methods for medical images. *Proceedings of the SPIE Visualization in Biomedical Computing Conference* **1992**; 1808: 336-346.

11. Puff DT, Pisano ED, Muller KE, Johnston RE, Hemminger BM, Burbeck CA, McLelland R, Pizer SM. A method for determination of optimal image enhancement for the detection of mammographic abnormalities. to appear in *The Journal of Digital Imaging.*

12. Smathers RL, Bush E, Drace J, Stevens M, Somer FG, Brown WB, Karras B. Mammographic microcalcifications: Detection with xerography, screen-film, and digitized film display. *Radiology* **1986**; 159: 673-677.